软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(4):1759−1776 [doi: 10.13328/j.cnki.jos.007460] [CSTR: 32375.14.jos.007460]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于分类检索的操作规划方法*

吴益露, 王瀚霖, 王利民

(南京大学 计算机学院, 江苏 南京 210023)

通信作者: 王利民, E-mail: lmwang@nju.edu.cn

摘　要: 聚焦于教学视频 (instructional video) 中的操作规划 (procedure planning) 问题, 探讨如何根据给定的开始和结束视觉状态, 在教学视频提供的动作空间中规划出一条将开始状态转变为结束状态的动作序列. 教学视频以记录和展示各种事件的操作过程为特点, 每个事件对应一组特定动作, 从而形成事件的动作空间. 多个事件的动作空间共同构成了教学视频的整体动作空间. 传统方法未能充分挖掘事件的语义信息, 过于依赖强化学习等复杂训练方法, 既增加了算法设计的复杂性, 又导致模型的可解释性较差. 针对这些问题, 结合教学视频的特点, 提出了一种基于分类检索的操作规划器 CPP (classification-retrieval-based procedure planner), 分阶段解决操作规划任务. 具体而言, 该方法首先通过视觉状态识别事件类别, 将动作空间限定在一个较小的子空间内, 显著降低规划的复杂性; 随后, 在该子空间中进行动作序列的规划. 此外, 提出了一种混合规划策略, 将动作序列的检索与预测相结合, 进一步提升了规划性能. 实验结果表明, 该方法在 3 个不同规模的教学视频数据集上均取得了显著效果, 为操作规划任务提供了一种简单而高效的基准方法.

关键词: 操作规划; 教学视频; 视觉推理

中图法分类号: TP18

中文引用格式: 吴益露, 王瀚霖, 王利民. 基于分类检索的操作规划方法. 软件学报, 2026, 37(4): 1759–1776. http://www.jos.org.cn/1000-9825/7460.htm

英文引用格式: Wu YL, Wang HL, Wang LM. Classification-retrieval-based Procedure Planning Method. Ruan Jian Xue Bao/Journal of Software, 2026, 37(4): 1759–1776 (in Chinese). http://www.jos.org.cn/1000-9825/7460.htm

Classification-retrieval-based Procedure Planning Method

WU Yi-Lu, WANG Han-Lin, WANG Li-Min

(School of Computer Science, Nanjing University, Nanjing 210023, China)

Abstract: This study focuses on the problem of procedure planning in instructional videos. Given the start and end observations, the task is to plan an action sequence that transforms the start state into the end state within the action space provided by the instructional videos. Instructional videos record and demonstrate the operational processes of various events. Each event includes a specific set of actions, forming the action space for that event. Therefore, the action space is composed of various subspaces corresponding to different events in instructional videos. Previous methods fail to effectively utilize the semantic information of events and overly rely on techniques such as reinforcement learning, resulting in complex training schemes and poorly explainable approaches. In contrast, this study considers the characteristics of instructional videos and proposes the classification-retrieval-based procedure planner (CPP), a pipeline that addresses procedure planning from coarse to fine. Specifically, the planner first identifies the event category based on the given observations, narrowing the action space to a smaller subspace. Then, action planning is performed within the selected subspace, which is significantly easier than planning in the entire action space. Moreover, this study introduces a hybrid planning method that combines retrieval and prediction approaches to generate the action sequence. The proposed method achieves competitive results on three popular procedure planning datasets of varying scales, establishing itself as a simple yet robust baseline for procedure planning.

Key words: procedure planning; instructional video; visual reasoning

* 基金项目: 科技创新 2030—“新一代人工智能”重大项目 (2022ZD0160900)
收稿时间: 2025-01-07; 修改时间: 2025-02-10; 采用时间: 2025-04-29; jos 在线出版时间: 2025-11-05
CNKI 网络首发时间: 2025-11-06

近年来, 视频理解领域[1−5]迅速发展, 在动作识别[6−8]、时序动作检测[9,10]和时空动作检测[11,12]等众多下游任务中取得了显著成果. 然而, 这些任务主要关注于视频中的短期语义信息, 而如何有效理解和利用长时语义信息仍然是亟待解决的挑战. 随着视频长度的不断增加, 更为复杂的视频理解任务逐渐受到关注, 例如时间维度上的推理[13]. 与此同时, 随着大量教学视频的提出[14−17], 针对教学视频的任务也日益丰富, 其中之一便是操作规划任务. 本文聚焦于教学视频中的操作规划问题, 该任务由 Chang 等人[18]首次提出, 是一个典型的长期视觉推理任务. 具体而言, 操作规划任务要求模型在给定开始和结束视觉状态的情况下, 规划出一个合理的动作序列, 以实现从开始状态到结束状态的转换, 如图 1 所示. 需要注意的是, 这些动作序列的动作均来源于由教学视频数据集定义的动作空间.

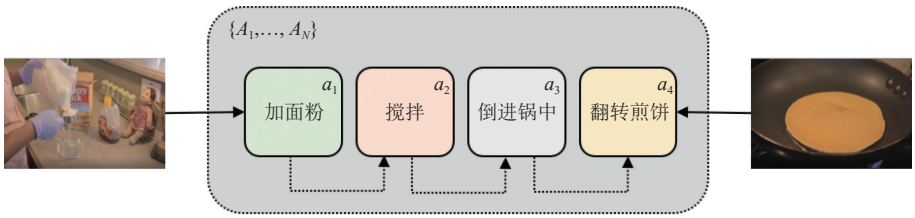


图 1 教学视频中的传统操作规划任务示意图

现有研究[19,20]在一些简单任务 (如在桌上堆叠积木) 的规划中表现出巨大潜力, 但在更复杂环境下的操作规划仍值得进一步研究. 尤其在教学视频中, 环境复杂且包含多种类型的动作, 动作持续时间各不相同, 这使得数据集的复杂性显著增加. 此外, 与简单规划任务相比, 教学视频中的操作规划仅能依赖开始和结束的视觉状态进行推理, 而中间状态的信息完全缺失, 这使得问题的设定更加困难. 数据集的复杂性和问题设定的难度共同构成了该任务的主要挑战.

目前已有许多方法[18,21−26]被提出, 在操作规划中进行了有意义的尝试. 我们在遵循以往工作的基础上, 提出了基于分类检索的操作规划器 (classification-retrieval-based retrieval procedure planner, CPP). 我们的方法主要基于以下 3 个动机.

(1) 现有的方法基于强化学习, 或使用复杂的损失函数来进行训练, 这些方法无论是在推理还是训练过程中都非常复杂. 在推理过程中, DDN[18]使用基于采样的前向规划[27]方法来规划动作序列; P^3IV[21]针对相同的开始和结束视觉信息, 需要生成 $K = 1500$ 个随机噪声样本, 再通过 Viterbi 后处理方法[28]进行平均. 在训练过程中, PDPP 等基于扩散模型的方法[23,29,30]需要经过大量的训练重建概率分布. 此外, 这些方法的规划结果只能在最终阶段看到, 导致解释性较差. 相比之下, 本文方法简单明了, 仅需使用简单的分类器进行训练. 在推理阶段, 我们提出了一种新的混合规划方法, 使规划过程中的问题能够被直观地观察到, 从而提升了解释性和易用性.

(2) 现有方法未能充分利用事件的语义信息, 而是在整个动作空间中直接进行操作规划. 这种方式存在明显问题. 如图 2 所示, 在方法 PlaTe[31]中, 有超过 1/3 的结果未能正确识别事件类别. 我们认为, 事件信息对操作规划至关重要, 因为动作空间由不同事件的子空间组成. 因此, 准确确定事件类别能够帮助缩小规划范围, 使规划更加高效和精确. 为此, 我们采用了一种直接有效的方法, 如图 3 所示. 我们提出了一种从粗到细的操作规划器, 首先通过分类确定事件类别, 接着在该事件类别对应的较小且更明确的动作空间中进行细粒度的规划, 从而显著提升了规划的效率与准确性.

(3) 进一步考虑了操作规划中的不确定性, 即对于相同的开始和结束动作, 可能存在多种合理的中间动作序列. 这种不确定性源于数据集中并非所有视频都严格按照固定顺序执行. 例如, 在“制作柠檬水”这一事件中, 开始和结束动作分别是“挤柠檬”和“搅拌混合物”. 在此过程中, 中间动作可能是“倒水”或“倒柠檬汁”, 两者均为合理选择. 单纯依赖语义信息进行规划难以准确判断中间动作. 然而, 如果结合视觉信息, 这一问题可能得到有效解决: 当场景中存在水时, 模型倾向于选择“倒水”; 而当场景中出现柠檬汁时, 模型则更倾向于选择“倒柠檬汁”. 基于此, 我们从语义和视觉两个层面共同规划中间动作, 提出了混合规划模块以更好地应对这一挑战.

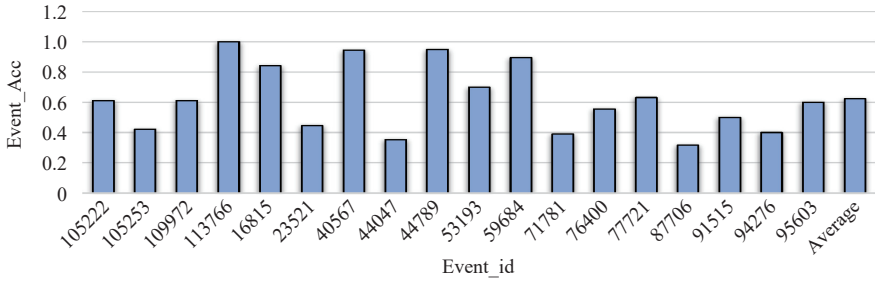


图 2 PlaTe 的事件准确率

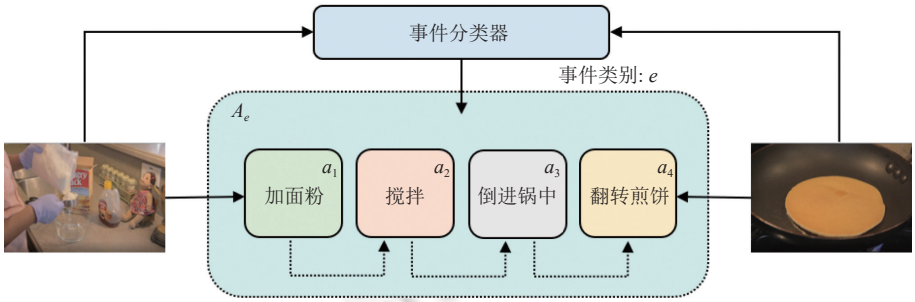


图 3 基于分类检索的操作规划任务示意图

鉴于教学视频具有事件级别和动作级别标签的特性, 我们首先通过一个简单的分类器确定当前的事件类别. 事件类别确定后, 将整个动作空间缩小到该事件对应的子空间, 在更小、更明确的子空间中进行细粒度规划. 随后, 我们设计了一种混合规划模块, 将检索方法和预测方法相结合, 进一步提升规划性能. 该方法设计简单明了, 充分利用了事件信息, 具有很强的可解释性, 并为后续优化提供了良好的基础. 在研究过程中, 我们还发现现有的操作规划评估指标存在不完全合理的问题, 目前的大多数方法的比较并不完全公平. 为此, 我们参考已有研究, 更新了更合理的操作规划评估指标. 在 COIN[17]、CrossTask[15]和 NIV[14]这 3 个主流教学视频数据集上的实验结果表明, 本文方法在各数据集上均表现出较强的竞争力, 尤其在 COIN 数据集上的操作规划指标取得了显著提升.

本文第 1 节介绍操作规划的相关工作与研究现状. 第 2 节具体介绍操作规划器, 包括事件分类模块和混合规划模块. 第 3 节通过实验验证本文方法的有效性, 给出在 3 个数据集上的实验结果, 并提供了详细的消融实验. 最后总结全文.

1 操作规划相关工作

教学在我们日常生活中至关重要, 教学的形式可以是文本、语音和视频. 其中视频形式的教学更加直观且易于理解, 因此近年来出现了越来越多的教学视频数据集[14−17,32−35]. 这些数据集的规模各不相同, 较小的数据集[14,33,34]只涉及少数事件, 如烹饪. 一些大规模数据集[16,17]则包含了来自不同领域的上百个事件类别. 这些教学视频通常包含事件级别的标注. 因此, 我们可以利用事件类别来解决教学视频的相关任务. 随着数据集数量的增加, 基于教学视频的研究也越来越多, 例如动作分割[36]、弱监督动作分割[37−39]、步骤定位[40,41]、视频字幕生成[32]以及视觉定位[42]等. 此外, 许多工作[43−46]旨在更好地理解教学视频的视觉信息. 本文研究的是长期视频推理任务, 即教学视频中的操作规划任务.

最初的操作规划任务是在一个非常简单的环境中规划序列. 随后, Chang 等人[18]提出了将操作规划迁移到更复杂的现实生活中, 因此操作规划自然扩展到了教学视频. 以往的规划方法主要分为: (1) 基于强化学习的方法[18,31,47], 它将规划任务视为马尔可夫决策问题[48], 并使用双分支结构分别拟合动作空间和特征空间中的真实分布. (2) 基于 Transformer 的方法[21,22,24,25,49], 基于 Transformer 架构[50]规划操作序列, 其中, P^3IV[21]利用了动作的文本信息进行弱监督; SCHEMA[49]使用大语言模型提供额外的中间状态描述来辅助训练; RAP[25]使用自回归 Trans-

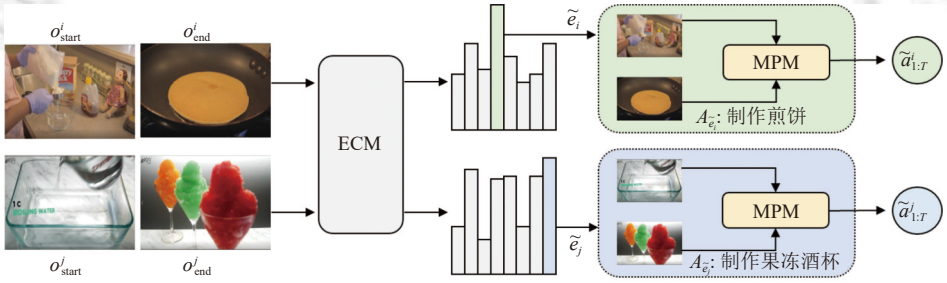
former 结构来生成可变长度的操作序列. (3) 基于 Diffusion 的方法[23,29,30], 将操作规划问题视为一个概率分布重建问题, 通过扩散模型重建期望的概率分布. 与以往使用复杂结构的方法相比, 我们选择使用更直接的基于分类检索的方法, 并充分利用教学视频的特性, 提出了一种新的从粗到细的操作规划器.

动作变化的推理或预测逐渐成为一个备受关注的研究方向. 在推理任务中, 模型需要根据已知的动作或视觉信息推断未见过的动作类别[51,52]、视觉信息的文本描述[53]以及其他形式[54,55]. 本文中的操作规划也是视觉推理任务之一. 目前, 动作预测和早期动作识别[56−59]是最受欢迎的视觉推理任务之一, 这些任务在 EPIC-KITCHENS 数据集[60]上得到了广泛研究. 这两类任务的核心在于推理短期的动作变化, 而操作规划则面向更复杂的场景, 需要推理长期的动作序列. 此外, 这些任务的输入通常仅包含开始的视觉状态, 而操作规划是一个目标驱动的任务, 需同时考虑开始和结束的视觉信息, 从而在规划中兼顾当前状态与目标状态的联系.

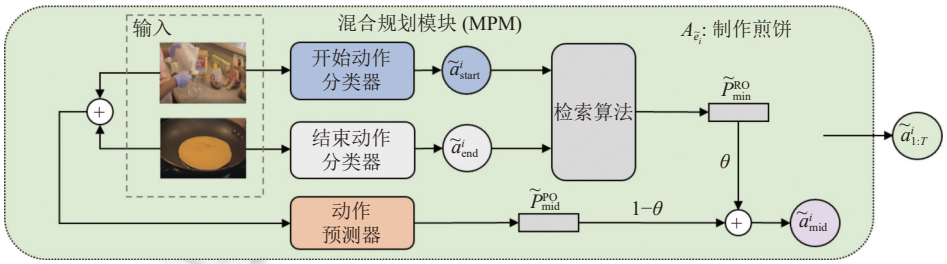
随着大规模视频文本数据集的出现, 例如 HowTo100M[16], 许多结合文本的多模态视频工作应运而生[41,43,61]. 主流的做法主要分为: (1) 采用对比损失[43,62], 将不同模态的信息进行融合; (2) 将基于掩码的目标函数作为多模态视频领域研究方向[3,63,64]; (3) 端到端的训练方法[16,43], 利用大量教学视频的训练, 旨在获得更好的视频和文本特征, 以便应用于下游任务的事件抽取. VideoCLIP[62]专注于长期的时间建模, 旨在捕捉丰富的上下文信息. 随后, Han 等人[65]研究了直接将上下文信息丰富的叙事文本与视频帧对齐的多模态视频模型. VINA[46]将视频中步骤定位作为最终目标, 而不是文本与视频对齐. 本文使用 MIL-NCE[43]模型来提取教学视频的视觉特征用于操作规划.

2 基于分类检索的操作规划器

本节将详细描述操作规划器. 首先提出操作规划的问题定义, 然后介绍事件分类模块 (event classification module, ECM), 该模块用于识别事件的类别. 最后介绍混合规划模块 (mixed planning module, MPM), 该模块包含 3 种不同的规划方法. 图 4 是所提方法的框架图, 图中展示了两个样本, 首先样本经过事件分类模块 (图 4(a) 中灰色模块) 确定事件类别. 根据确定的事件类别将动作空间缩小到特定的子空间中, 在图 4(a) 中样本 i 对应的子空间用绿色模块表示; 样本 j 对应的子空间用蓝色模块表示. 随后, 样本在缩小的子空间中通过混合规划模块进行规划, 对应于图 4(a) 中的黄色模块. 图 4(b) 中以样本 i 为例, 展示了混合规划模块的详细结构, 首先通过开始和结束动作分类器确定开始和结束的动作, 再结合检索和预测两种方法确定中间的动作, 从而确定整个序列.



(a) 基于两个样本下的操作规划器 (CPP) 框架图



(b) MPM 混合规划模块, 对应于 (a) 中黄色框

图 4 基于分类检索的操作规划器 (CPP) 框架图和混合规划模块示意图

2.1 问题设定

对于给定的视觉信息 O_{start} 和 O_{end}, 我们的模型需要根据给定的动作空间 A 规划出一个动作序列 $\tilde{\pi}$, 其中动作空间由数据集提供. 动作序列表示为:

$$\tilde{\pi} = \{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_T\},\ \tilde{a}_i \in A \tag{1}$$

其中, T 表示序列中包含多少个动作, $\tilde{a}_i$ 为动作 (标记了 ~ 的符号是模型推理出的结果, 未标记的符号表示真实的结果). 教学视频数据集中包含了多种事件的视频, 所以数据集的动作空间就是由多种事件的子动作空间构成, 所以动作空间可以表示为:

$$A = \{A_1 \cup A_2 \cup \ldots \cup A_N\} \tag{2}$$

其中, N 是数据集中事件的个数, 即子动作空间的个数. 之前的工作在完整的动作空间 A 中进行操作规划, 这样使得规划的解空间过大, 所以规划的难度较大. 为了缩小规划的动作空间, 我们首先根据 O_{start} 和 O_{end} 确定当前的事件类别 $\tilde{e}$, 然后将动作空间缩小为 $\tilde{e}$ 的子空间 $A_{\tilde{e}}$. 此时问题转变为:

$$\tilde{\pi} = \{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_T\},\ \tilde{a}_i \in A_{\tilde{e}} \tag{3}$$

在确定了事件类别后, 我们将操作规划视为一个分类问题, 将动作序列分为 3 个部分:

$$\tilde{\pi} = \{\tilde{a}_{\text{start}}, \tilde{a}_{\text{mid}}, \tilde{a}_{\text{end}}\},\ \tilde{a}_{\text{start}}, \tilde{a}_{\text{mid}}, \tilde{a}_{\text{end}} \in A_{\tilde{e}} \tag{4}$$

其中, $\tilde{a}_{\text{start}}$, $\tilde{a}_{\text{end}}$ 分别为开始和结束的动作. $\tilde{a}_{\text{mid}}$ 为中间动作, 包含了 $T-2$ 个动作 (去除了开始和结束的动作). 我们首先根据 O_{start} 和 O_{end} 确定 $\tilde{a}_{\text{start}}$ 和 $\tilde{a}_{\text{end}}$, 然后使用检索和预测相结合的方法来确定最优的 $\tilde{a}_{\text{mid}}$. 第 2.3 节将详细介绍不同的规划方法.

2.2 事件分类模块

本节介绍事件分类模块. 我们通过将 O_{start} 和 O_{end} 输入到 ECM 中得到当前的事件类别. 具体来说, 我们将开始和结束的视觉特征在特征维度连接起来, 而后输入到一个事件分类器中. 因为输入的是抽取好的特征, 所以不需要特别复杂的模型, 我们选择了一个简单的 4 层的多层感知器 (multilayer perceptron, MLP) 结构作为 ECM. 多层感知器的输出是所有事件的概率分布 $\widetilde{\mathbf{p}}_e$, 对应于真值的独热编码 (one-hot encoding) $\mathbf{e}$, 我们用交叉熵损失函数来训练这个分类器, 损失函数如下:

$$L_{\text{ECM}} = -\mathbf{e}\log\widetilde{\mathbf{p}}_e \tag{5}$$

根据获得的事件概率, 我们选择概率最高的事件作为规划子空间, $\tilde{e} = \arg\max(\widetilde{\mathbf{p}}_e)$.

2.3 混合规划模块

本节详细描述了混合规划模块. 首先介绍开始动作分类器 (action classifiers) 和结束动作分类器, 接着详细介绍规划方法. 我们展示了 3 种规划方法: 仅检索方法 (retrieval-only method)、仅预测方法 (prediction-only method) 和混合规划方法 (mixed-planning method). 在最终结果展示中, 我们选择了混合规划方法, 详细的消融实验将在第 3.6 节展示.

2.3.1 动作分类器

经过 ECM 过程后, 规划任务被缩小到一个新的动作空间 $A_{\tilde{e}}$. 可规划的动作数量从 n 变为 $n_{\tilde{e}}$. 然后, 在此新空间中确定 $\tilde{a}_{\text{start}}$ 和 $\tilde{a}_{\text{end}}$. 由于开始和结束观测是动作片段的一部分, 我们使用一个简单的 3 层 MLP 作为分类器进行识别. 但与常规的动作识别任务相比, 我们的分类更为困难, 因为只能看到部分动作信息. 如图 4 所示, 我们将 O_{start} 和 O_{end} 分别输入对应的开始动作和结束动作分类器, 获得开始和结束动作的概率, 分别为 $\widetilde{\mathbf{p}}_{\text{start}}$ 和 $\widetilde{\mathbf{p}}_{\text{end}}$. 开始和结束的真实标签的独热编码表示为 $\mathbf{a}_{\text{start}}$ 和 $\mathbf{a}_{\text{end}}$, 这里同样使用交叉熵损失函数:

$$L_{\text{SC}} = -\mathbf{a}_{\text{start}}\log\widetilde{\mathbf{p}}_{\text{start}},\ L_{\text{EC}} = -\mathbf{a}_{\text{end}}\log\widetilde{\mathbf{p}}_{\text{end}} \tag{6}$$

根据获得的概率, 我们选择概率最高的动作作为分类结果:

$$\tilde{a}_{\text{start}} = \arg\max(\widetilde{\mathbf{p}}_{\text{start}}),\ \tilde{a}_{\text{end}} = \arg\max(\widetilde{\mathbf{p}}_{\text{end}}) \tag{7}$$

2.3.2 仅检索方法

在获得开始和结束预测类别后, 仅检索方法在训练集中检索 $\tilde{a}_{\text{mid}}$, 训练集可以表示为以下形式:

$$\left\{a_{\text{start}}^k, a_{\text{mid}}^k, a_{\text{end}}^k\right\}_{k=1}^K \tag{8}$$

其中, K 表示长度为 T 的样本数量. 我们固定相同的 $\tilde{a}_{\text{start}}$ 和 $\tilde{a}_{\text{end}}$, 在训练集检索可能的中间动作, 并计算动作出现的概率 $\tilde{\mathbf{p}}_{\text{mid}}^{\text{RO}}$, RO 表示仅检索方法. 我们将检索算法表示为 R, 如算法 1 所示.

算法 1. 检索算法 R.

输入: $\tilde{a}_{\text{start}}$, $\tilde{a}_{\text{end}}$, $\mathcal{A}_{\tilde{e}}$, $\pi = \{a_{\text{start}}^k, a_{\text{mid}}^k, a_{\text{end}}^k\}_{k=1}^K$; ▷ 输入: 预测的开始和结束动作, 子动作空间, 训练集

输出: $\tilde{\mathbf{p}}_{\text{mid}}$. ▷ 输出: 中间动作的概率分布

```
count ← [0]^K                                   ▷ 初始化计数数组为 0, 大小为 K
match ← 0                                       ▷ 初始化匹配标志变量为 0
for k = 1 to K do                               ▷ 遍历训练集样本
    if ã_start = a_start^k and ã_end = a_end^k then
        count[a_mid^k] ← count[a_mid^k] + 1
        match ← 1                               ▷ 标记为已找到完全匹配的序列
    end if
end for
if match = 0 then                               ▷ 如果没有找到完全匹配的开始和结束动作
    for k = 1 to K do                           ▷ 遍历所有样本, 查找部分匹配
        if ã_start = a_start^k or ã_end = a_end^k then
            count[a_mid^k] ← count[a_mid^k] + 1
        end if
    end for
end if
p̃_mid ← Softmax(count)                          ▷ 将计数数组通过 Softmax 转换为概率分布
return p̃_mid                                    ▷ 返回中间动作的概率分布
```

整个算法过程看似非常简单, 但存在两个问题: (1) 在检索过程中, 成对的 $\tilde{a}_{\text{start}}$ 和 $\tilde{a}_{\text{end}}$ 可能在训练集中无法匹配. 也就是说, 我们可能无法在训练集中找到相同的 $(\tilde{a}_{\text{start}}, \tilde{a}_{\text{end}})$ 对. 出现此问题有两种情况, 一种是由于 $\tilde{a}_{\text{start}}$ 或 $\tilde{a}_{\text{end}}$ 的错误, 导致该对无法匹配. 另一种情况是, 即使 $\tilde{a}_{\text{start}}$ 和 $\tilde{a}_{\text{end}}$ 都正确, 由于训练集和测试集之间的差异, 仍然无法匹配, 即测试集中的一些样本在训练集中并未出现. 为了解决这个问题, 如算法 1 所示, 我们将检索问题从匹配开始和结束动作简化为仅匹配开始动作或仅匹配结束动作. (2) 根据 $\tilde{\mathbf{p}}_{\text{mid}}^{\text{RO}}$, 选择概率最高的一个作为 $\tilde{a}_{\text{mid}}$, 但可能会有多个结果具有相同的概率. 在仅检索的方法中, 我们使用最直接的随机选择方法来确定结果, 也就是考虑到这种情况的出现, 后面结合了视觉信息来预测中间动作类别. 我们首先尝试仅预测方法, 然后尝试将两种方法结合.

2.3.3 仅预测方法

我们使用仅预测方法代替检索来直接获得动作概率 $\tilde{\mathbf{p}}_{\text{mid}}^{\text{PO}}$, PO 表示仅预测方法. 动作预测器的输入是开始和结束特征, 如图 4(b) 所示, 它也是一个简单的多层感知器, 同样使用交叉熵损失进行监督, 这里我们不展示具体的计算公式. 我们将在第 3.5 节中展示实验结果.

2.3.4 混合规划方法

最后, 我们选择结合检索和预测的方法. 如图 4(b) 所示, 我们通过对两种概率 $\tilde{\mathbf{p}}_{\text{mid}}^{\text{RO}}$ 和 $\tilde{\mathbf{p}}_{\text{mid}}^{\text{PO}}$ 进行加权求和来结合

这两种方法. 最终的 $\widetilde{\mathbf{p}}_{\text{mid}}$ 计算如下:

$$\widetilde{\mathbf{p}}_{\text{mid}} = \theta\widetilde{\mathbf{p}}_{\text{mid}}^{\text{RO}} + (1-\theta)\widetilde{\mathbf{p}}_{\text{mid}}^{\text{PO}} \tag{9}$$

其中, θ 是一个超参数, 最终的中间动作结果为 $\tilde{a}_{\text{mid}} = \arg\max(\widetilde{\mathbf{p}}_{\text{mid}})$.

2.4 训练与测试

在训练时, 我们分别训练事件分类器、动作分类器和动作预测器, 每个模型均是一个简单的 MLP. 开始和结束分类器使用相似的结构, 但并不共享参数. 训练动作分类器时, 我们使用数据增强, 即将视频中每个动作的观测作为训练样本. 在推理时, 我们将批次大小设置为 1, 通过事件分类器确定事件类别, 选择相应的开始和结束分类器以获得开始和结束动作类别, 通过混合规划模块获得完整的序列.

3 实验分析

本节详细介绍实验设置, 具体包括数据集介绍、评价指标和样本提取. 我们在 3 个不同的数据集上评估本文方法, 并与之前的方法进行比较. 此外, 我们验证了本文方法中每部分的效果. 随后, 讨论操作规划中的不确定性. 最后, 展示方法的可视化结果.

3.1 数据集介绍

我们遵循之前的工作[21−26], 在 3 个主流的教学视频数据集上评估我们的方法, 数据集分别是: COIN[17]、CrossTask[15]和 NIV[14].

COIN 是规模庞大的教学视频数据集, 包含 11 827 个视频、12 个领域、180 个事件和 778 个动作. 我们遵循以往的方法[21−26], 按照 70%/30% 的比例对数据集进行划分.

CrossTask 数据集, 包含 2 750 个视频、18 个事件和 133 个动作, 其事件主要涉及烹饪任务, 如“制作塔可沙拉”“烤牛排”. CrossTask 提供了官方划分方案, 训练集和测试集分别包含 2 390 个和 360 个视频. 尽管部分方法[31]使用了官方划分方案, 其他方法[21−26]均采用了 70%/30% 的划分策略创建训练/测试集. 为了更公平地比较, 我们在此采用 70%/30% 的数据划分.

NIV 是一个相对较小的数据集, 仅包含 150 个视频、5 个事件和 48 个动作. 由于 NIV 未提供官方数据划分方案, 我们同样以 70%/30% 的比例对数据集进行划分.

除了规模差异外, CrossTask 和 NIV 每个视频的动作数量超过 COIN. COIN 更为标准, 大多数动作遵循标准顺序, 而 CrossTask 和 NIV 中的动作可能会重复多次或不按标准顺序进行. 因此, 我们的混合规划模块在 COIN 上的表现优于其他两个数据集. 详细结果将在第 3.5 节中展示.

3.2 评价指标

为了与之前的工作[21−26]保持一致, 我们使用了 3 个已有的指标来评估性能. (1) 平均交并比 (*mIoU*), 这是一个最不严格的指标. 该指标要求模型输出正确的动作, 但不要求动作的顺序正确. 这里, 我们对 *mIoU* 进行特别解释. 根据公开的代码, 之前的方法[21,22]在整个批次上计算 *mIoU*, 即将整个批次的动作添加到集合中. 我们认为这种做法不够严谨, 因为批次大小会影响最终结果, 随着批次大小的增加, *mIoU* 会增加. 因此, 我们在单个样本上计算 IoU. 所以与之前的工作相比, 我们的 *mIoU* 会略低, 但指标更严谨. (2) 平均准确率 (Acc) 该指标考虑每一步动作的准确性, 不要求序列完全正确, 只关注每个单独的动作是否正确. (3) 成功率 (*SR*) 要求整个预测动作序列与真实的动作序列完全一致, 这是最严格的指标. (4) 事件准确率 (EventAcc) 考虑预测序列是否能识别当前事件类别, 当序列中有多于一半的动作成功识别事件类别, 我们就认为这个序列成功识别. (5) 开始准确率 (SAcc) 和结束准确率 (EAcc) 分别计算开始和结束动作的平均准确率. 中间准确率 (MAcc) 表示中间动作的平均准确率. 我们提出的新指标 EventAcc、SAcc、EAcc 与 MAcc 基于原始指标, 可以细化操作规划任务, 更准确地评估模型.

3.3 样本提取

每个教学视频都标注了描述人类动作的动作标签和边界. 我们遵循现有的方法[21−26], 将每个视频视为图像序列 $I_{1:L}$, 其中包含 M 个动作片段, 每个片段带有动作标签 $a_{1:M}$ 和时间边界 $(ts_{1:M}, te_{1:M})$. 我们希望获取每个动作的开

始状态点和结束状态点, 即 os_i 和 oe_i. 对于第 i 个动作片段, 我们选择开始时间点附近的图像 $I_{ts_{i-\sigma},ts_{i+\sigma}}$ 作为 os_i, 并选择结束时间点附近的图像 $I_{te_{i-\sigma},te_{i+\sigma}}$ 作为 oe_i, 其中 $\sigma=1$. 对于操作规划, 需要长度为 T 的动作序列作为样本, 但是大多数视频中的动作数量并不等于 T. 对于少于 T 个动作的视频, 我们会通过重复的方法补齐动作序列. 对于多于 T 个动作的视频, 我们使用滑动窗口 (长度为 T) 来剪裁视频, 以此来考虑所有的动作序列. 所以最终每个样本包含 T 个动作, os_0 作为开始状态, oe_{T-1} 作为结束状态.

3.4 实现细节

本文规划器输入的是视频特征. 对于 3 个数据集, 我们按照 P^{3}IV[21]方法使用 S3D[43]提取特征, 该模型使用 HowTo100M[16]数据集进行预训练, 以实现视频-文本联合嵌入. 我们通过枚举每个事件包含的动作, 并基于事件类别划分不同的子空间来构建动作空间. 具体来说, CrossTask 包含 18 个子空间, 每个子空间平均有 7.4 个动作; NIV 包含 5 个子空间, 每个事件平均有 9.5 个动作; 对于包含 12 个领域和 180 个事件的 COIN, 我们在事件级别划分子空间, 因此包含 180 个子空间, 每个子空间平均包含 4.3 个动作.

3.5 实验结果

我们比较了关于指导性视频中的操作规划的工作. (1) 随机策略 (random policy): 该基线随机从动作空间中选择动作序列, 代表性能的下限. (2) 基于检索的方法 (retrieval): 该基线通过在特征空间中最小化起点和终点对的距离, 匹配训练集中最近的结果. 我们的方法与该基线类似, 不同之处在于该方法在特征空间中进行检索, 而我们的方法则先进行分类, 再在动作空间中进行检索. (3) DDN[18]: 该方法首次提出将操作规划任务扩展到教学视频. DDN 使用双分支模型来学习状态-动作转移的双动态. (4) P^{3}IV[21]: 该方法使用弱语言监督进行操作规划. P^{3}IV 是一个基于 Transformer[50]的单分支模型, 并配备了一个记忆模块. (5) E3P[22]: 该方法通过将事件信息编码到序列建模过程中, 从环境状态中规划动作序列, 并依据预测事件规划后续动作. 不同于该方法, 本文方法直接利用事件信息以缩小动作空间, 实验结果表明其效果更优. (6) PDPP[23]: 该方法将操作规划任务视为拟合分布的过程, 通过扩散模型重建分布. (7) Skip-plan[24]: 将操作规划视为一个数学链问题, 该模型通过绕过不可靠的中间动作, 将较长的链条分解为多个较短的子链条. (8) RAP[25]: 该模型克服了传统方法的限制, 能够处理固定和可变的动作序列, 并通过弱语言监督和可学习的检索式记忆组件, 提升了序列规划的能力. (9) KEPP[26]: 该方法通过构造知识图谱的方法来增强操作规划中的过程知识.

我们首先在当前规模最大的教学视频数据集 COIN 上对方法进行了测试, 结果如表 1 所示, 其中加粗的数据表示最优结果, 带下划线的数据表示次优结果. 与之前基于大规模模型且训练方式更为复杂的方法相比, 本文方法在多个指标上均表现出显著提升. 当 $T=3$ 时, *SR* 指标提升 6.39%; 当 $T=4$ 时, *SR* 指标提升 7.38%. 同样地, 准确率也有着大幅改进, 分别在 $T=3$ 和 $T=4$ 时提高了 5.79% 和 6.16%. 需要特别指出的是, 我们对 *mIoU* 的计算进行了标准化: 与之前方法在整个批次上计算 *mIoU* 不同, 我们在单个样本上计算 *mIoU*. 因此, 与之前方法相比, 我们的 *mIoU* 结果可能相对偏低.

表 1 COIN 上预测范围为 $T=3$, $T=4$ 的操作规划结果 (%)

方法	$T=3$			$T=4$		
	SR	Acc	*mIoU*	*SR*	Acc	*mIoU*
Random	<0.01	<0.01	2.47	<0.01	<0.01	2.32
Retrieval	4.38	17.40	32.06	2.71	14.29	36.97
DDN[18]	13.90	20.19	64.78	11.13	17.71	68.06
P^{3}IV[21]	15.40	21.67	76.31	11.32	18.85	70.53
E3P[22]	19.57	31.42	84.95	13.59	26.72	84.72
PDPP[23]	21.33	45.62	51.82	14.41	<u>44.10</u>	51.39
Skip-plan[24]	23.65	47.12	78.44	16.04	43.19	77.07
RAP[25]	<u>24.47</u>	<u>47.59</u>	85.24	<u>16.76</u>	42.11	84.64
KEPP[26]	20.25	39.87	51.72	15.63	39.53	53.27
Ours	**30.86**	**53.38**	60.40	**24.14**	**50.26**	61.87

COIN 数据集包含超过 1 万个视频, 是涵盖操作种类最多、事件范围最广的教学视频数据集. 本文方法在消耗最少训练资源的情况下, 取得了最优的结果, 这充分证明了方法的有效性. 我们认为能够在 COIN 上取得优异表现的原因主要有以下两点. 首先, COIN 包含 180 个事件类别, 动作空间庞大. 相比于以往在整个动作空间中直接规划的方法, 我们首先通过确定事件类别缩小动作空间, 大幅降低了问题的复杂性. 根据表 2 中展示的更多结果, COIN 的事件分类准确率达到 80%, 较之前的工作有显著提升. 其次, COIN 是一个相对规范的教学视频数据集, 这也使得我们的检索方法能够显著提升规划的成功率.

表 2 COIN 上更多操作规划结果 (%)

T	EventAcc	SAcc	EAcc	MAcc	*SR*	Acc	*mIoU*
3	81.33	58.03	57.01	45.10	30.86	53.38	60.40
4	80.50	58.30	56.79	42.98	24.14	50.26	61.87
5	79.43	58.24	56.85	44.60	21.35	49.78	62.01

我们同样在较小的数据集 CrossTask 上测试了操作规划结果. 如表 3 所示, 本文方法同样取得了具有竞争力的表现. 虽然未能达到最优结果, 但当 $T=3$ 和 $T=4$ 时, *SR* 指标均为次优, 这对于我们仅需极少训练资源的方法来说, 是一个不错的结果. 值得注意的是, PDPP 的结果标注了*, 这是为了确保公平比较, 我们采用了与标准设定相同的结果, 而非 PDPP 论文中的原始结果. 我们认为 CrossTask 上的提升不如 COIN, 主要原因在于, 与 COIN 相比, CrossTask 的操作序列中存在大量动作重复和动作缺失, 导致不确定性更高. 更多关于不确定性分析的讨论可参见第 3.7 节. 尽管如此, 我们的混合规划模块仍然有效, 确保了方法的结果足以超越大多数之前的方法. 表 4 展示了在 CrossTask 上的更多操作规划结果. 随着 *T* 值的增加, 事件分类的准确率保持稳定, 但由于检索难度加大, 整体正确率出现了明显下降.

表 3 CrossTask 上预测范围为 $T=3$, $T=4$ 的操作规划结果 (%)

方法	$T=3$			$T=4$		
	SR	Acc	*mIoU*	*SR*	Acc	*mIoU*
Random	<0.01	0.94	1.66	<0.01	0.83	1.66
Retrieval	8.05	23.30	32.06	3.95	22.22	35.97
DDN[18]	12.18	31.29	47.48	5.97	27.10	48.46
P^3IV[21]	23.34	49.96	73.89	13.40	44.16	70.01
E3P[22]	26.40	53.02	74.05	16.49	48.00	70.16
PDPP*[23]	26.38	55.62	59.34	18.69	52.44	62.38
Skip-plan[24]	28.85	<u>61.18</u>	74.98	15.56	55.64	70.30
RAP[25]	29.23	**62.35**	75.63	16.96	<u>55.77</u>	71.49
KEPP[26]	**33.38**	60.79	63.89	**21.02**	**56.08**	64.15
Ours	<u>31.68</u>	59.02	61.09	<u>19.36</u>	53.82	60.55

表 4 CrossTask 更多操作规划结果 (%)

T	EventAcc	SAcc	EAcc	MAcc	*SR*	Acc	*mIoU*
3	90.66	68.93	61.32	46.82	31.68	59.02	61.09
4	91.52	69.64	62.06	41.80	19.36	53.82	60.55
5	92.03	71.02	62.70	37.85	10.79	49.46	59.84
6	92.76	71.89	63.58	37.53	6.87	47.60	59.12

我们在较小的数据集 NIV 上对方法进行了评估. 如表 5 所示, 本文方法在 $T=3$ 时取得了最高的 *SR* 指标 (26.61%), 并在 $T=4$ 时实现了最佳的 Acc 指标 (48.93%), 较基准提升约 7%. 由于 NIV 数据集规模较小, 实验中存在一定的过拟合现象, 导致结果波动较大, 没有任何方法能够在所有指标上全面领先. 我们还在更长时间区间上评

估了 NIV 的表现, 结果如表 6 所示. 可以看出, 我们的模型即使在较小的数据集上, 也展现出了较强的长期规划能力. 此外, 由于 NIV 数据集仅包含 5 个任务, 我们的方法在事件分类上的准确率几乎接近 100%.

表 5　NIV 上预测范围为 $T=3$, $T=4$ 的操作规划结果 (%)

方法	$T=3$			$T=4$		
	SR	Acc	*mIoU*	*SR*	Acc	*mIoU*
Random	2.21	4.07	6.09	1.12	2.73	5.84
DDN[18]	18.41	32.54	56.56	15.97	27.09	53.84
P^{3}IV[21]	24.68	49.01	74.29	20.14	38.36	67.29
E3P[22]	26.05	**51.24**	75.81	21.37	41.96	74.90
KEPP[26]	24.44	43.46	86.67	**22.71**	41.59	91.49
Ours	**26.61**	48.79	55.40	20.00	**48.93**	55.81

表 6　NIV 更多操作规划结果 (%)

T	EventAcc	SAcc	EAcc	MAcc	*SR*	Acc	*mIoU*
3	99.60	50.00	48.79	47.58	26.61	48.79	55.40
4	100.00	49.52	51.91	47.14	20.00	48.93	55.81
5	100.00	45.67	50.29	42.20	11.56	44.51	57.31
6	99.27	46.72	45.26	40.15	5.84	42.09	56.88

3.6　消融实验

在本节中, 我们评估了各个模块的性能. 为了验证事件分类模块的有效性, 我们移除了 ECM 模块, 直接在整个动作空间中对开始和结束动作进行分类, 并且检索预测中间动作. 根据表 7 所示, 移除 ECM 模块后, 在 COIN 和 CrossTask 数据集上的事件分类准确率分别下降了约 5% 和 6%. NIV 由于动作空间非常小, 所以事件分类准确率没有发生变化. 剩下的所有指标在这 3 个数据集上都有大幅度的下降, 开始和结束动作的准确率下降了超过 5%. 由于开始和结束的准确率直接影响我们中间动作的检索结果, 最终的 *SR* 指标和 Acc 指标也显著下降. 尤其是 NIV 的 *SR* 指标下降了约 12%. 这些结果表明, 首先通过 ECM 缩小规划的动作空间是非常必要的.

表 7　事件分类模块消融实验 (%)

数据集	方法	EventAcc	SAcc	EAcc	MAcc	*SR*	Acc	*mIoU*
COIN	CPP without ECM	76.63	50.51	51.12	40.83	21.00	47.49	53.46
	CPP	**81.33**	**58.03**	**57.01**	**45.10**	**30.86**	**53.38**	**60.40**
CrossTask	CPP without ECM	84.77	63.46	57.47	43.37	26.02	54.77	56.19
	CPP	**90.66**	**68.93**	**61.32**	**46.82**	**31.68**	**59.02**	**61.09**
NIV	CPP without ECM	99.60	45.97	39.52	39.11	14.92	41.53	47.39
	CPP	**99.60**	**50.00**	**48.79**	**47.58**	**26.61**	**48.79**	**55.40**

根据第 2.3 节中描述的规划方法, 我们将混合规划方法替换为仅检索方法 (RO) 和仅预测方法 (PO), 分别对其效果进行评估. 如表 8 所示, 仅检索方法 (RO) 已能够实现较为优异的规划性能, 特别是在 $T=4$ 时, 在 CrossTask 数据集上取得了最高的 *SR* 指标 19.44%, 在 NIV 数据集上取得了最高的 *mIoU* 指标 55.94%, 这充分验证了检索方法的有效性. 相较之下, 仅预测方法 (PO) 的 *SR* 指标出现了显著下降, 但却在 Acc 指标上表现突出. 在 COIN 和 NIV 数据集上, 仅预测方法均取得了最佳的 Acc 指标.

为了分析其中的原因, 我们从概率论的视角进行解释. 对于仅预测方法, 开始动作、中间动作和结束动作的概率被假设为条件独立, 因此最终的 *SR* 指标实际上是这 3 个部分联合概率的结果, 即:

$$P(SR)=P(\tilde{a}_{\text{start}})P(\tilde{a}_{\text{mid}})P(\tilde{a}_{\text{end}}) \tag{10}$$

表 8 混合规划模块消融实验 (%)

数据集	方法	$T=3$			$T=4$		
		SR	Acc	*mIoU*	*SR*	Acc	*mIoU*
COIN	CPP (RO)	30.63	53.03	60.10	23.96	49.64	61.47
	CPP (PO)	28.50	**54.96**	58.99	20.70	**52.36**	60.61
	CPP (Ours)	**30.86**	53.38	**60.40**	**24.14**	50.26	**61.87**
CrossTask	CPP (RO)	31.59	58.37	60.52	**19.44**	52.68	59.93
	CPP (PO)	27.32	58.47	59.27	14.55	53.82	59.39
	CPP (Ours)	**31.68**	**59.02**	**61.09**	19.36	**53.82**	**60.55**
NIV	CPP (RO)	26.21	47.58	54.59	19.05	46.55	**55.94**
	CPP (PO)	18.15	**49.33**	53.38	15.71	**49.88**	54.81
	CPP (Ours)	**26.61**	48.79	**55.40**	**20.00**	48.93	55.81

这种独立性假设导致中间动作的准确率较低, 从而影响 *SR* 指标. 而在仅检索方法中, 中间动作的检索结果基于开始和结束动作的准确率, 体现为条件概率关系, 即:

$$P(\tilde{a}_{\text{mid}}|\tilde{a}_{\text{start}},\tilde{a}_{\text{end}}) \tag{11}$$

根据实验结果, 当开始和结束动作完全正确时, 中间动作的检索准确率分别为 COIN 数据集约 70%、CrossTask 数据集约 60%、NIV 数据集约 80%, 显著高于仅预测方法中中间动作的独立预测准确率. 因此, 仅检索方法在 *SR* 指标上的表现更为优异. 综上, 我们结合仅检索方法和仅预测方法的优势, 设计了混合规划模块, 以平衡 *SR* 指标和 Acc 指标的表现, 进一步提升了方法的整体性能.

我们接下来介绍如何确定混合规划模块中的超参数 θ. 由于操作规划任务仅划分了训练集和测试集, 为了确定超参数, 我们从训练集中随机抽取 20% 的数据作为验证集进行调优. 由于其他方法用全部训练集来训练, 我们划分验证集仅用于确定超参数. 对于 COIN、CrossTask 和 NIV, 最终的 θ 值设置为 0.8、0.7 和 0.9.

表 8 展示了仅检索方法 (RO) 和仅预测方法 (PO) 的实验结果, 分别对应于 $\theta=1$ 和 $\theta=0$. 此外为了进一步验证超参数的合理性, 表 9 展示了在 COIN 测试集上, 不同 θ 值的消融实验结果, $\theta=0.8$ 的这一行数据为我们最终的结果. 实验结果表明, 随着检索模块的权重变大, *SR* 指标逐渐升高, 但去除掉预测模块 ($\theta=1$), *SR* 指标会下降. 当检索模块的权重降低, 预测模块的权重增大时, Acc 指标升高, 当仅使用预测模块时 ($\theta=0$), Acc 指标达到最大值, 这与我们前文分析的结论一致.

表 9 COIN 上不同 θ 消融实验 (%)

方法	$T=3$			$T=4$		
	SR	Acc	*mIoU*	*SR*	Acc	*mIoU*
CPP ($\theta=0$)	28.50	**54.96**	58.99	20.70	**52.36**	60.61
CPP ($\theta=0.6$)	30.70	53.74	60.40	23.86	50.67	61.80
CPP ($\theta=0.7$)	30.79	53.55	**60.50**	23.92	50.40	61.77
CPP ($\theta=0.8$)	**30.86**	53.38	60.40	24.14	50.26	**61.87**
CPP ($\theta=0.9$)	30.86	53.32	60.30	**24.18**	50.09	61.70
CPP ($\theta=1$)	30.63	53.03	60.10	23.96	49.64	61.47

3.7 不确定性讨论

不确定性是过程规划任务中的一个重要特征. 这种不确定性体现为, 对于相同的开始和结束动作, 可能存在多种合理的中间序列. 例如, 在制作一份复杂料理的过程中, 从清洗食材到最终摆盘, 可能有不同的步骤排列方式, 且这些序列在语义上都是合理的. 然而, 在最初提出该问题时, 任务仅要求模型生成一个最佳解并进行评估, 且相关数据集仅提供了一个标准答案, 未覆盖所有可能的合理序列. 基于此设定, 我们的方法在当前阶段主要生成一个最佳解作为规划结果. 值得强调的是, 我们的方法具有天然的扩展性, 可以适应多解场景. 具体而言, 我们将检索到的

多种可能解视为合理解. 在这一框架下, 模型可以生成整个解分布而非单一解. 我们进一步参考 P^{3}IV 的思路, 将模型生成的规划分布与数据集的标准分布进行对比, 从而全面评估模型对多种解的支持能力. 我们采用了两个经典的分布比较指标: KL 散度 (Kullback-Leibler divergence, KL-Div) 和负对数似然 (negative log-likelihood, NLL). 此外, 为了进一步验证模型的多解能力, 我们引入了召回率 (ModeRec) 和精确度 (ModePrec) 两个指标. ModeRec 衡量模型对标准答案的覆盖能力, 即在生成的解分布中是否包含数据集中的标准解. 而 ModePrec 则评价模型生成解的合理性和可行性, 即生成解在测试数据中的准确性.

为了构建标准分布, 本文检索了测试集中所有满足给定开始和目标状态的动作序列, 作为参考的标准规划分布. 如表 10 所示, 概率方法更好地匹配了标准分布, 取得了更好的 KL-Div 和 NLL 指标. 这是符合预期的结果, 表明本文检索出的合理的解是在准确的解空间中, 也表明了我们缩小动作空间办法的有效性. 概率方法在 ModeRec 指标上表现出明显优势, 这表明模型生成的解分布能够更好地覆盖标准解. 然而, 由于概率方法允许更大的解空间, 这也导致了 ModePrec 指标的下降. 解空间的扩展虽然增加了生成合理序列的机会, 但也不可避免地引入了一些错误解, 降低了解的整体准确性. 值得注意的是, 我们仅在 CrossTask 数据集上进行了不确定性实验. 这是因为, 与 COIN 和 NIV 数据集相比, CrossTask 的不确定性更高, 即在相同的开始和结束状态下存在更多的合理解序列.

表 10 CrossTask 不确定性结果 (%)

指标	方法	$T=3$	$T=4$	$T=5$	$T=6$
KL-Div↓	Ours-deterministic	3.65	3.84	4.05	4.09
	Ours-probabilistic	**2.96**	**2.62**	**2.41**	**2.26**
NLL↓	Ours-deterministic	4.22	4.80	5.28	5.51
	Ours-probabilistic	**3.54**	**3.58**	**3.65**	**3.69**
ModePrec↑	Ours-deterministic	**46.26**	**41.81**	**34.80**	**25.37**
	Ours-probabilistic	42.75	33.64	24.19	17.14
ModeRec↑	Ours-deterministic	30.20	18.53	10.39	6.49
	Ours-probabilistic	**45.21**	**38.51**	**29.37**	**21.86**

3.8 *mIoU* 指标的讨论

在测试过程中, 不同的测试批次大小会对 *mIoU* 指标产生显著影响. 根据第 3.2 节中的说明, 我们在批次大小为 1 时对 *mIoU* 进行了计算, 以确保结果的公平性和可比性. 分别从理论和实验的角度, 说明批次大小对 *mIoU* 指标的影响. 首先, *mIoU* 指标的计算公式如下:

$$mIoU = \frac{1}{N}\sum_{i}^{N}\frac{|\tilde{\pi}\cap\pi|}{|\tilde{\pi}\cup\pi|} \tag{12}$$

其中, $\tilde{\pi}$ 是预测序列的动作集合, π 是真实序列的动作集合, N 为所有的样本数. 当在整个批次上计算 *mIoU* 指标时, $\tilde{\pi}$ 是整个批次中所有预测序列的动作的集合, π 是整个批次中所有真实序列的动作的集合, N 为批次数, 这样会导致 *mIoU* 指标被高估. 因为同一个批次中的错误预测可能会相互抵消, 极端情况下可能会出现两个样本预测结果互换的情况, 即一个样本预测了另一个样本的正确序列, 这会使得 *mIoU* 指标看起来很高, 但实际上每个样本的预测结果可能并不好. 并且我们也通过实验证明, 随着批次大小的增加, *mIoU* 值也逐渐提高, 当批次大小增加至 20 时, *mIoU* 可达到约 90% (CrossTask 中可达到 93.04%). 所以对于操作规划任务, 我们认为在推理阶段对每个样本独立计算 *mIoU* 更能真实评估模型的实际效果, 从而避免因批次处理而导致的评估偏差.

4 可视化

我们给出了规划操作序列的可视化示例, 其中绿色框表示正确的动作, 红色框表示错误的动作. 如图 5 所示, 对于相同的开始和结束观测, 仅检索方法会以相同的概率得到两个中间动作结果, 因为这些示例均在训练集中出现过. 具体来说, 训练集中包含了“将煎饼从锅中拿出并翻面”的示例. 然而, 当我们使用混合规划方法并引入视觉

信息时, 模型能够准确判断煎饼是否仍在锅中, 从而生成更为准确的操作序列. 这一结果表明, 结合视觉信息可以显著提升模型在处理复杂场景时的判断能力.

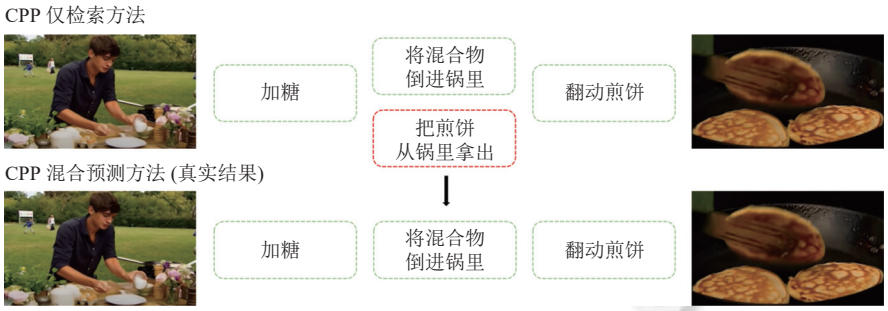


图 5 混合规划模块有效性可视化

在图 6 中, 我们展示了操作规划序列长度为 3、4 的可视化结果, 并且与其他方法 (PDPP) 的结果进行对比. 每个样本包含开始与结束视觉状态以及规划结果. 我们 CPP 的规划结果第 1 行表示动作序列, 第 2 行是分类的事件类别. 由于 PDPP 没有事件分类模块, 所以只有一行动作序列. 其中绿色框表示正确结果, 红色框表示错误结果. 两个样本中我们的方法准确预测了完整的动作序列, 而 PDPP 在样本 1 中最后一个动作发生错误, 在样本 2 中对于事件判断错误, 导致大部分动作都不正确. 我们还提供了序列长度为 3 时的失败的规划结果. 图 7 样本 1 展示了我们混合规划模块出错的示例, 其中开始和结束动作是正确的, 但模型规划了错误的中间动作. 图 7 样本 2 中, 我们展示了即使没有得到正确的开始或结束动作, 也有可能规划出正确的中间动作, 但这种情况非常少见. 图 7 样本 3 描述了整个序列完全错误的示例, 由于事件类别错误, 模型在错误的动作子空间中进行了规划, 导致整个序列错误.

图 6 正确样本的可视化结果

图 7 错误样本的可视化结果

5 总 结

本文提出了一种基于简单分类检索的操作规划方法. 不同于以往的方法, 该方法通过事件分类器采用粗到细的思路, 将操作规划的动作空间有效缩小至特定的子空间. 随后, 整个规划序列在子空间中通过混合规划方法生成. 本文在 3 个主流数据集上对该方法进行了评估, 结果表明其在短期和长期规划任务中均表现出色. 尽管事件分类模块显著提升了性能, 但事件分类错误问题仍未完全解决. 如图 7 所示, 当事件分类出错时, 整个序列规划结果也会随之出错. 然而, 与缩小动作空间所带来的显著改进相比, 这类错误的影响相对较小. 未来工作可进一步探索解决该问题的有效方法. 此外, 尽管检索方法的准确性较高, 其计算开销仍需优化, 减少开销将是进一步研究的重点方向. 长期视频理解是一个极具潜力的研究方向, 希望未来能够提出一个通用模型, 用于教学视频的长期理解,

从而满足不同下游任务的需求.

References

[1] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

[2] Wang XL, Girshick R, Gupta A, He KM. Non-local neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803. [doi: 10.1109/CVPR.2018.00813]

[3] Tong Z, Song YB, Wang J, Wang LM. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022.

[4] Wang LM, Huang BK, Zhao ZY, Tong Z, He YN, Wang Y, Wang YL, Qiao Y. VideoMAE V2: Scaling video masked autoencoders with dual masking. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14549–14560. [doi: 10.1109/CVPR52729.2023.01398]

[5] Li KC, He YN, Wang Y, Li YZ, Wang WH, Luo P, Wang YL, Wang LM, Qiao Y. VideoChat: Chat-centric video understanding. arXiv:2305.06355, 2023.

[6] Zhou BL, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 831–846. [doi: 10.1007/978-3-030-01246-5_49]

[7] Feichtenhofer C, Fan HQ, Malik J, He KM. SlowFast networks for video recognition. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6201–6210. [doi: 10.1109/ICCV.2019.00630]

[8] Wang LM, Xiong YJ, Wang Z, Qiao Y, Lin DH, Tang XO, van Gool L. Temporal segment networks for action recognition in videos. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740–2755. [doi: 10.1109/TPAMI.2018.2868668]

[9] Lin TW, Liu X, Li X, Ding ER, Wen SL. BMN: Boundary-matching network for temporal action proposal generation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3888–3897. [doi: 10.1109/ICCV.2019.00399]

[10] Lin TW, Zhao X, Su HS, Wang CJ, Yang M. BSN: Boundary sensitive network for temporal action proposal generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–21. [doi: 10.1007/978-3-030-01225-0_1]

[11] Song L, Zhang SW, Yu G, Sun HB. TACNet: Transition-aware context network for spatio-temporal action detection. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11979–11987. [doi: 10.1109/CVPR.2019.01226]

[12] Yang XT, Yang XD, Liu MY, Xiao FY, Davis LS, Kautz J. STEP: Spatio-temporal progressive learning for video action detection. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 264–272. [doi: 10.1109/CVPR.2019.00035]

[13] Farha YA, Ke QH, Schiele B, Gall J. Long-term anticipation of activities with cycle consistency. In: Proc. of the 42nd German Conf. on Pattern Recognition. Tübingen: Springer, 2020. 159–173. [doi: 10.1007/978-3-030-71278-5_12]

[14] Alayrac JB, Bojanowski P, Agrawal N, Sivic J, Laptev I, Lacoste-Julien S. Unsupervised learning from narrated instruction videos. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4575–4583. [doi: 10.1109/CVPR.2016.495]

[15] Zhukov D, Alayrac JB, Cinbis RG, Fouhey D, Laptev I, Sivic J. Cross-task weakly supervised learning from instructional videos. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3532–3540. [doi: 10.1109/CVPR.2019.00365]

[16] Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2630–2640. [doi: 10.1109/ICCV.2019.00272]

[17] Tang YS, Lu JW, Zhou J. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(9): 3138–3153. [doi: 10.1109/TPAMI.2020.2980824]

[18] Chang CY, Huang DA, Xu DF, Adeli E, Li FF, Niebles JC. Procedure planning in instructional videos. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 334–350. [doi: 10.1007/978-3-030-58621-8_20]

[19] Finn C, Tan XY, Duan Y, Darrell T, Levine S, Abbeel P. Deep spatial autoencoders for visuomotor learning. In: Proc. of the 2016 IEEE Int'l Conf. on Robotics and Automation. Stockholm: IEEE, 2016. 512–519. [doi: 10.1109/ICRA.2016.7487173]

[20] Finn C, Levine S. Deep visual foresight for planning robot motion. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation. Singapore: IEEE, 2017. 2786–2793. [doi: 10.1109/ICRA.2017.7989324]

[21] Zhao H, Hadji I, Dvornik N, Derpanis KG, Wildes RP, Jepson AD. P³IV: Probabilistic procedure planning from instructional videos with

weak supervision. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2928–2938. [doi: 10.1109/CVPR52688.2022.00295]

[22] Wang AL, Lin KY, Du JR, Meng JK, Zheng WS. Event-guided procedure planning from instructional videos with text supervision. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 13519–13529. [doi: 10.1109/ICCV51070.2023.01248]

[23] Wang HL, Wu YL, Guo S, Wang LM. PDPP: Projected diffusion for procedure planning in instructional videos. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14836–14845. [doi: 10.1109/CVPR52729.2023.01425]

[24] Li ZH, Geng WJ, Li MH, Chen L, Tang YS, Lu JW, Zhou J. Skip-plan: Procedure planning in instructional videos via condensed action space learning. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 10263–10272. [doi: 10.1109/ICCV51070.2023.00945]

[25] Zare A, Niu YL, Ayyubi H, Chang SF. RAP: Retrieval-augmented planner for adaptive procedure planning in instructional videos. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 410–426. [doi: 10.1007/978-3-031-72980-5_24]

[26] Nagasinghe KRY, Zhou HL, Gunawardhana M, Min MR, Harari D, Khan MH. Why not use your textbook? Knowledge-enhanced procedure planning of instructional videos. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 18816–18826. [doi: 10.1109/CVPR52733.2024.01780]

[27] Ghallab M, Nau D, Traverso P. Automated Planning: Theory and Practice. San Francisco: Morgan Kaufmann Publishers Inc., 2004.

[28] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. on Information Theory, 1967, 13(2): 260–269. [doi: 10.1109/TIT.1967.1054010]

[29] Fang F, Liu Y, Koksal A, Xu QL, Lim JH. Masked diffusion with task-awareness for procedure planning in instructional videos. arXiv:2309.07409, 2023.

[30] Shi L, Bürkner PC, Bulling A. ActionDiffusion: An action-aware diffusion model for procedure planning in instructional videos. In: Proc. of the 2025 IEEE/CVF Winter Conf. on Applications of Computer Vision. Tucson: IEEE, 2025. 8816–8825. [doi: 10.1109/WACV61041.2025.00854]

[31] Sun JK, Huang DA, Lu B, Liu YH, Zhou BL, Garg A. PlaTe: Visually-grounded planning with Transformers in procedural tasks. IEEE Robotics and Automation Letters, 2022, 7(2): 4924–4930. [doi: 10.1109/LRA.2022.3150855]

[32] Das P, Xu CL, Doell RF, Corso JJ. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2634–2641. [doi: 10.1109/CVPR.2013.340]

[33] Stein S, McKenna SJ. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proc. of the 2013 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing. Zurich: ACM, 2013. 729–738. [doi: 10.1145/2493432.2493482]

[34] Kuehne H, Arslan A, Serre T. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 780–787. [doi: 10.1109/CVPR.2014.105]

[35] Zhou LW, Xu CL, Corso J. Towards automatic learning of procedures from Web instructional videos. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 7590–7598. [doi: 10.1609/AAAI.V32I1.12342]

[36] Rohrbach M, Amin S, Andriluka M, Schiele B. A database for fine grained activity detection of cooking activities. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 1194–1201. [doi: 10.1109/CVPR.2012.6247801]

[37] Kuehne H, Richard A, Gall J. Weakly supervised learning of actions from transcripts. Computer Vision and Image Understanding, 2017, 163: 78–89. [doi: 10.1016/J.CVIU.2017.06.004]

[38] Richard A, Kuehne H, Gall J. Action sets: Weakly supervised action segmentation without ordering constraints. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5987–5996. [doi: 10.1109/CVPR.2018.00627]

[39] Lu ZJ, Elhamifar E. Set-supervised action learning in procedural task videos via pairwise order consistency. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19871–19881. [doi: 10.1109/CVPR52688.2022.01928]

[40] Elhamifar E, Huynh D. Self-supervised multi-task procedure learning from instructional videos. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 557–573. [doi: 10.1007/978-3-030-58520-4_33]

[41] Cao M, Yang TY, Weng JW, Zhang C, Wang J, Zou YX. LocVTP: Video-text pre-training for temporal localization. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 38–56. [doi: 10.1007/978-3-031-19809-0_3]

[42] Huang DA, Buch S, Dery L, Garg A, Li FF, Niebles JC. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018.

5948–5957. [doi: 10.1109/CVPR.2018.00623]

[43] Miech A, Alayrac JB, Smaira L, Laptev I, Sivic J, Zisserman A. End-to-end learning of visual representations from uncurated instructional videos. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9876–9886. [doi: 10.1109/CVPR42600.2020.00990]

[44] Zhou HL, Martín-Martín R, Kapadia M, Savarese S, Niebles JC. Procedure-aware pretraining for instructional video understanding. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 10727–10738. [doi: 10.1109/CVPR52729.2023.01033]

[45] Zhong YW, Yu LC, Bai Y, Li SW, Yan XT, Li Y. Learning procedure-aware video representation from instructional videos and their narrations. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14825–14835. [doi: 10.1109/CVPR52729.2023.01424]

[46] Mavroudi E, Afouras T, Torresani L. Learning to ground instructional articles in videos through narrations. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 15155–15167. [doi: 10.1109/ICCV51070.2023.01395]

[47] Bi J, Luo JB, Xu CL. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 15591–15600. [doi: 10.1109/ICCV48922.2021.01532]

[48] Bellman R. A Markovian decision process. Indiana University Mathematics Journal, 1957, 6(4): 679–684. [doi: 10.1512/iumj.1957.6.56038]

[49] Niu YL, Guo WL, Chen L, Lin XD, Chang SF. SCHEMA: State CHangEs MAtter for procedure planning in instructional videos. arXiv:2403.01599, 2024.

[50] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[51] Farha YA, Richard A, Gall J. When will you do what?—Anticipating temporal occurrences of activities. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5343–5352. [doi: 10.1109/CVPR.2018.00560]

[52] Furnari A, Farinella G. What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6251–6260. [doi: 10.1109/ICCV.2019.00635]

[53] Liang C, Wang WG, Zhou TF, Yang Y. Visual Abductive Reasoning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15544–15554. [doi: 10.1109/CVPR52688.2022.01512]

[54] Tan C, Yeo CK, Tan C, Fernando B. Inferring past human actions in homes with abductive reasoning. arXiv:2210.13984, 2022.

[55] Chi HG, Lee K, Agarwal N, Xu Y, Ramani K, Choi C. AdamsFormer for spatial action localization in the future. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17885–17895. [doi: 10.1109/CVPR52729.2023.01715]

[56] Sener F, Singhania D, Yao A. Temporal aggregate representations for long-range video understanding. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 154–171. [doi: 10.1007/978-3-030-58517-4_10]

[57] Girdhar R, Grauman K. Anticipative video Transformer. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 13485–13495. [doi: 10.1109/ICCV48922.2021.01325]

[58] Stergiou A, Damen D. The wisdom of crowds: Temporal progressive attention for early action prediction. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14709–14719. [doi: 10.1109/CVPR52729.2023.01413]

[59] Zhong ZY, Schneider D, Voit M, Stiefelhagen R, Beyerer J. Anticipative feature fusion Transformer for multi-modal action anticipation. In: Proc. of the 2023 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2023. 6057–6066. [doi: 10.1109/WACV56688.2023.00601]

[60] Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W, Wray M. Scaling egocentric vision: The EPIC-KITCHENS dataset. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 753–771. [doi: 10.1007/978-3-030-01225-0_44]

[61] Ko D, Choi J, Ko J, Noh S, On KW, Kim ES, Kim HJ. Video-text representation learning via differentiable weak temporal alignment. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5006–5015. [doi: 10.1109/CVPR52688.2022.00496]

[62] Xu H, Ghosh G, Huang PY, Okhonko D, Aghajanyan A, Metze F, Zettlemoyer L, Feichtenhofer C. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 6787–6800.

[63] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proc. of

the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR, 2020.

[64] Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: A joint model for video and language representation learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7463–7472. [doi: 10.1109/ICCV.2019.00756]

[65] Han TD, Xie WD, Zisserman A. Temporal alignment networks for long-term video. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2896–2906. [doi: 10.1109/CVPR52688.2022.00292]

作者简介

吴益露, 博士, 主要研究领域为教学视频理解, 技能学习.

王瀚霖, 硕士, 主要研究领域为深度学习, 视频理解, 视频生成.

王利民, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 深度学习, 视频理解.